Preliminary Experiments with Geo-Filtering Predicates for Geographic IR

Jochen L. Leidner[1,2,3]
[1] Linguit GmbH, Friedensstraße 10,
76887 Bad Bergzabern, Germany.
⟨leidner@linguit.com⟩

[2] University of the Saarland, FR 7.4 – Speech Signal Processing,
Building 17.1, Office 0.17, 66041 Saarbrücken, Germany.

[3] University of Edinburgh, School of Informatics,
2 Buccleuch Place, Edinburgh EH8 9LW, Scotland UK.

Abstract

This paper describes a set of experiments for monolingual English retrieval at GEO-CLEF 2005. We evaluate a technique for spatial retrieval based on named entity tagging, toponym resolution, and re-ranking by means of geographic filtering. To this end, we present a series of systematic experiments in the Vector Space paradigm. We investigate plain bag-of-word versus a kind of phrasal retrieval, the potential of meronymic query expansion as a recall-enhancing device, and compare three alternative geo-spatial filtering techniques based on spatial clipping. We evaluate these on 25 monolingual English queries.

Our prelimiary results show that always choosing toponym referents based on a simple "maximum population" heuristic to approximate the salience of a referent fails to outperform TF*IDF baselines with the GEO-CLEF 2005 dataset when combined with three geo-filtering predicates. Conservative geo-filtering outperforms more aggresive predicates. The evidence further seems to suggest that query expansion with WordNet meronyms is not effective in combination with the method described.

A cursory post-hoc analysis indicates that responsible factors for the low performance include sparseness of available population data, gaps in the gazetteer that associates Minumum Bounding Rectangles with geo-terms in the query, and the composition of the GEO-CLEF 2005 dataset itself.

1 Introduction

Since all human activity relates to places, a large number of information needs also contain a geographic or otherwise spatial aspect. People want to know about the *nearest* restaurant, about the outcome of the match football match *in Manchester*, or about how many died in a flood in *in Thailand*. Traditional IR however, does not accomodate this spatial aspect enough: place names or geographic expressions are merely treated as strings, just like other query terms. This paper presents a general technique to accomodate geographic space in IR, and presents an evaluation of a particular instance of it carried out withn CLEF 2005.

1.1 The CLEF Evaluation

The Cross-Language Evaluation Forum (CLEF)[1] is an initiative funded by the European Union, and part of the DELOS Network of Excellence for Digital Libraries (EU FW-6). It aims to study competing IR methods across a variety of languages and tasks in annual international evaluations.

[1] http://www.clef-campaign.org/

```
<top>
  <num> GC001 </num>
  <orignum> C084 </orignum>
  <EN-title>Shark Attacks off Australia and California</EN-title>
  <EN-desc> Documents will report any information relating to shark
    attacks on humans. </EN-desc>
  <EN-narr> Identify instances where a human was attacked by a shark,
    including where the attack took place and the circumstances
    surrounding the attack. Only documents concerning specific attacks
    are relevant; unconfirmed shark attacks or suspected bites are not
    relevant.
  </EN-narr>
  <!-- NOTE: This topic has added tags for GeoCLEF -->
  <EN-concept> Shark attacks </EN-concept>
  <EN-spatialrelation>near</EN-spatialrelation>
  <EN-location> Australia </EN-location>
  <EN-location> California </EN-location>
</top>
```

Figure 1: The Anatomy of a GEO-CLEF Query (GC001).

1.2 The GEO-CLEF Track

In 2005, CLEF for the first time incorporated a track to study the performance of information retrieval strategies that take into account the notion of geographic space. This GEO-CLEF track, organized by the universities Berkeley and Sheffield, have the objective:

> “to compare methods of query translation, query expansion, translation of geographical references, use of text and spatial retrieval methods separately or combined, retrieval models and indexing methods.” (from the CLEF homepage)

In the first GEO-CLEF track, the languages English (monolingual), German (monolingual and cross-lingual), Portuguese and Spanish (cross-lingual) were offered. GEO-CLEF queries contain a geographic aspect (cf. Figure 1) that express spatial relevance contraints. Figure 2 lists the topic titles of the 25 English test queries used for GEO-CLEF in 2005. Each of the queries is run against a corpus which is a sub-collection of 56,472 *Glasgow Herald* documents and 113,005 documents from the *LA Times*. Obviously, this gives the corpus a distinct Scottish-Californian geographic bias, as we shall se later.

Paper plan. The remainder of this paper is organized as follows. Section 2 describes the method used to enhance IR with spatial knowledge. We present the experimental results obtained in the GEO-CLEF 2005 evaluation in Section 3, and summarize and conclude in Section 4 with some suggestion for future work.

2 Method

This section describes the method used in this study. Figure 3 shows the experimental setup. There are four essential processing steps. A document retrieval engine (IR) retrieves a set of documents relevant to the queries and groups them in a ranked list. A named entity tagging phase (NERC) then identifies all toponyms. Afterwards a toponym resolution (TR) module looks up all candidate referents for each toponym (i.e, the locations that the place name may be referring to) and tries to disambiguate the toponyms based on a heuristic. If successful, it also assigns the latitude/longitude of the centroid of the location to the resolved toponym. For each document-query pair a geo-filtering module (CLIP) then discards all locations outside a Minimum Bounding Rectangle (MBR) that is the denotation of the spatial expression in the query. Finally, based on a so-called geo-filtering predicate, it is decided whether or not the document under investigation

GC001	Shark Attacks off Australia and California
GC002	Vegetable Exporters of Europe
GC003	AI in Latin America
GC004	Actions against the fur industry in Europe and the U.S.A.
GC005	Japanese Rice Imports
GC006	Oil Accidents and Birds in Europe
GC007	Trade Unions in Europe
GC008	Milk Consumption in Europe
GC009	Child Labor in Asia
GC010	Flooding in Holland and Germany
GC011	Roman cities in the UK and Germany
GC012	Cathedrals in Europe
GC013	Visits of the American president to Germany
GC014	Environmentally hazardous Incidents in the North Sea
GC015	Consequences of the genocide in Rwanda
GC016	Oil prospecting and ecological problems in Siberia
GC017	American Troops in Sarajevo, Bosnia-Herzegovina
GC018	Walking holidays in Scotland
GC019	Golf tournaments in Europe
GC020	Wind power in the Scottish Islands
GC021	Sea rescue in North Sea
GC022	Restored buildings in Southern Scotland
GC023	Murders and violence in South-West Scotland
GC024	Factors influencing tourist industry in Scottish Highlands
GC025	Environmental concerns in and around the Scottish Trossachs

Figure 2: Test Queries Used in GEO-CLEF 2005.

Freq.	Toponym	Freq.	Toponym
18,452	**Scotland**	**4,140**	**Glasgow**
13,556	U.S.	4,347	China
9,013	**Los Angeles**	4,235	Washington
9,007	United States	4,013	England
7,893	California	3,985	America
7,458	Japan	3,817	Bosnia
7,294	Europe	3,548	France
6,985	**Orange County**	**3,388**	**Valley**
5,476	Britain	3,273	Russia
5,391	Metro	3,067	New York
4,686	Germany	**2,964**	**Edinburgh**
4,438	City	2,919	Mexico
4,400	London	2,782	**Southern California**

Table 1: List of the Most Frequent Toponyms in the GEO-CLEF corpus. Toponyms in bold type are artifacts of the Glasgow/California bias of the corpus.

is to be discarded, propagating up subsequent documents in the ranking. We now describe each phase in more detail.

2.1 Document Retrieval (IR)

The document retrieval engine provides access to the indexed GEO-CLEF collection. No stop-word filtering or stemming was used at index time, and index access is case-insensitive. The IR engine is used to retrieve the top 1,000 documents for each evaluation query from the collection using the Vector Space Model with a plain vanilla TF*IDF ranking function:

$$score(d,q) = \sum_{\forall t in q} tf(t,d)\ idf(t)\ lengthNorm(t,d) \qquad (1)$$

([GH05] p. 78 f.). We used the *Lucene* 1.4.3 search API for vector space retrieval [Cut05, GH05].[2] We use *Lucene*'s document analysis functionality for English text without modification.

2.2 Named Entity Tagging (NERC)

For named entity tagging, we use a Maximum Entropy classifier trained on MUC-7 data [CC03]. Tagging 1,000 retrieved document is a very expensive procedure; in a production system, this step would probably be carried out at indexing time. Therefore, the retrieved documents are actually pooled across runs to speed up the processing.

2.3 Toponym Resolution (TR)

For looking up the candidate referents, we use the large-scale gazetteer described in [Lei] as primary gazetteer, supplemented by the *World Gazetteer*[3] for population information (as secondary gazetteer). The algorithm used to resolve toponyms to referents works as follows: first, we look up the potential referents with associated latitude/longitude from the primary gazetteer. Then we look up population information for candidate referents from the secondary gazetteer. In order to relate the population entries from the *World Gazetteer* to corresponding entries of the main gazetteer, we defined a custom equality operator ($\doteq$) between two candidate referents for a toponym T_{R_i} such that $R_1 \doteq R_2$ holds iff there is a string equality

[2] There is also *CLucene*, a faster C implementation [van05], but at the time of writing it is slightly less mature than the Java implementation.

[3] http://worldgazetteer.com/

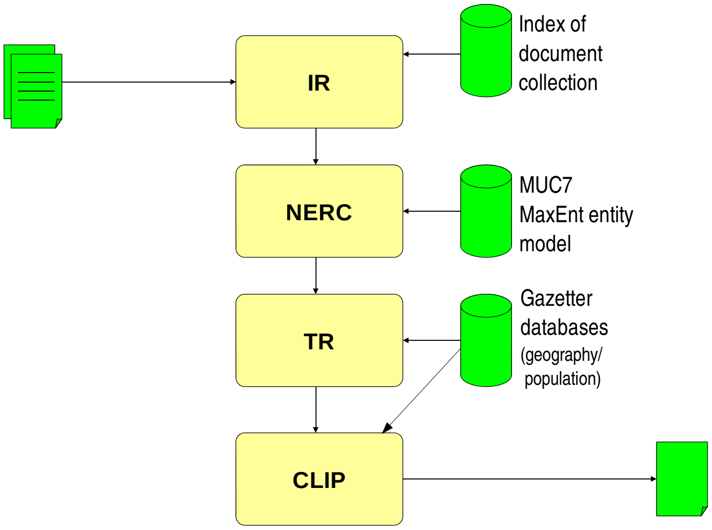

Index of document collection
IR
NERC
MUC7 MaxEnt entity model
TR
Gazetter databases (geography/ population)
CLIP

Figure 3: Experimental Setup Used in this Study.

between their toponyms ($T_{R_1} = T_{R_2}$) and the latitude and longitude of the candidate referents are in the same 1-degree grid (i.e., if and only if $[R_{1_{\text{lat}}}] = [R_{2_{\text{lat}}}] \wedge [R_{1_{\text{long}}}] = [R_{2_{\text{long}}}]$). If there is no population information available, the toponym remains unresolved (partial algorithm). If there is exactly one population entry, the toponym is resolved to that entry. If more than one candidate has population information available, the referent with the largest population is selected. Figure 4 shows the algorithm at work. In the example at the top a case is shown where only population information (prefixed by an asterisk) for one referent is available. This is used as evidence for that referent being the most salient candidate, and consequently it is selected. In the example shown at the bottom, population numbers for two candidate referents are available; the place with the larger number of inhabitant is selected. Note that the coordinates in the two gazetteers need to be rounded in order to ensure the matching of corresponding entries is successful.

Out of the 41,360 toponym types[4], population information was available in the World Gazetteer for *some* (i.e., more than zero) candidate referents only for 4,085 toponyms. This means that using only the population heuristics, the upper bound for system recall is $R = 9.88\%$, and for F-Score $F_{\beta=1} = 9.41\%$, assuming perfect resolution precision. Once the best toponym resolution strategy has become more apparent, it would be preferable to compute it offline and store the results in a spatially-enabled database management system such as *PostgreSQL* [vOV91].

2.4 Geographic Filtering (CLIP)

In principle, we can introduce a notion of geographic relevance in an existing approach to information retrieval in at least two ways: in a *ranking-based approach* the relevance metric gets directly modified to take locations in the document and query into account, i.e. instead of, say, using

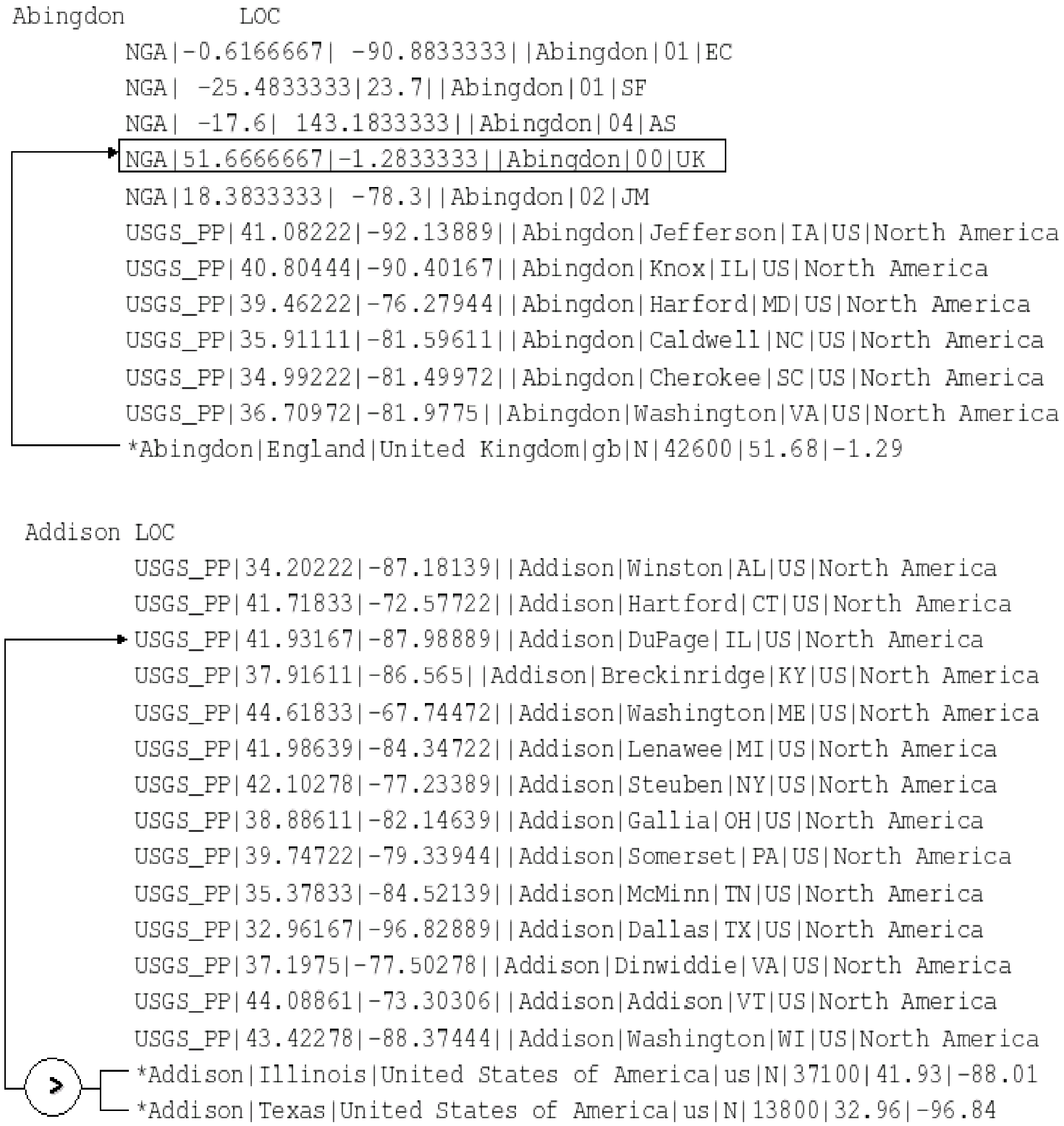
$$\text{Score}(d) = \text{TFIDF}(d) \tag{2}$$

we need a geographic relevance measure, $\text{Geo-Score}(d)$, which we may combine with our term-based score using linear interpolation:

$$\text{Score}'(d) = \lambda\,\text{Score}(d) + (1-\lambda)\,\text{Geo-Score}(d). \tag{3}$$

Alternatively, we may use a *filtering-based approach* such as the one attempted here, in which we apply traditional IR and then identify locations by means of toponym recognition and toponym resolution. We

[4] This number excludes "toponyms" that start with a digit (false positives caused by the NE tagger).



```
Abingdon        LOC
        NGA|-0.6166667| -90.8833333||Abingdon|01|EC
        NGA| -25.4833333|23.7||Abingdon|01|SF
        NGA| -17.6| 143.1833333||Abingdon|04|AS
        NGA|51.6666667|-1.2833333||Abingdon|00|UK
        NGA|18.3833333| -78.3||Abingdon|02|JM
        USGS_PP|41.08222|-92.13889||Abingdon|Jefferson|IA|US|North America
        USGS_PP|40.80444|-90.40167||Abingdon|Knox|IL|US|North America
        USGS_PP|39.46222|-76.27944||Abingdon|Harford|MD|US|North America
        USGS_PP|35.91111|-81.59611||Abingdon|Caldwell|NC|US|North America
        USGS_PP|34.99222|-81.49972||Abingdon|Cherokee|SC|US|North America
        USGS_PP|36.70972|-81.9775||Abingdon|Washington|VA|US|North America
        *Abingdon|England|United Kingdom|gb|N|42600|51.68|-1.29

 Addison LOC
        USGS_PP|34.20222|-87.18139||Addison|Winston|AL|US|North America
        USGS_PP|41.71833|-72.57722||Addison|Hartford|CT|US|North America
        USGS_PP|41.93167|-87.98889||Addison|DuPage|IL|US|North America
        USGS_PP|37.91611|-86.565||Addison|Breckinridge|KY|US|North America
        USGS_PP|44.61833|-67.74472||Addison|Washington|ME|US|North America
        USGS_PP|41.98639|-84.34722||Addison|Lenawee|MI|US|North America
        USGS_PP|42.10278|-77.23389||Addison|Steuben|NY|US|North America
        USGS_PP|38.88611|-82.14639||Addison|Gallia|OH|US|North America
        USGS_PP|39.74722|-79.33944||Addison|Somerset|PA|US|North America
        USGS_PP|35.37833|-84.52139||Addison|McMinn|TN|US|North America
        USGS_PP|32.96167|-96.82889||Addison|Dallas|TX|US|North America
        USGS_PP|37.1975|-77.50278||Addison|Dinwiddie|VA|US|North America
        USGS_PP|44.08861|-73.30306||Addison|Addison|VT|US|North America
        USGS_PP|43.42278|-88.37444||Addison|Washington|WI|US|North America
    >   *Addison|Illinois|United States of America|us|N|37100|41.93|-88.01
        *Addison|Texas|United States of America|us|N|13800|32.96|-96.84
```

Figure 4: Toponym Resolution Using the Maximum-Population Heuristic.

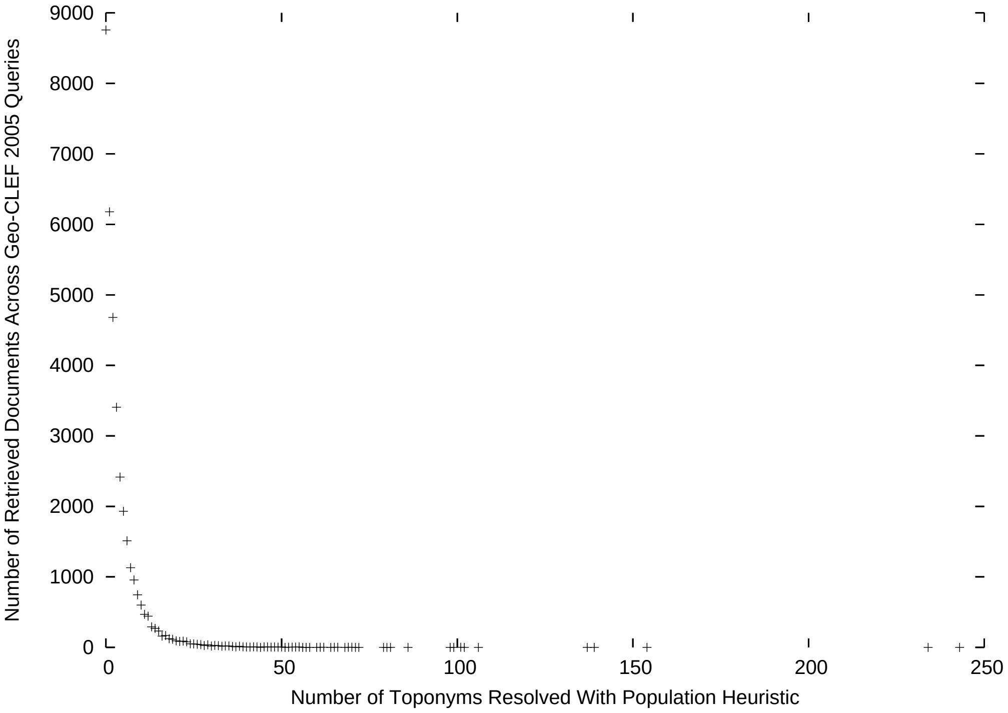


Figure 5: Number of Toponyms Resolved in Document Against Number of Documents.

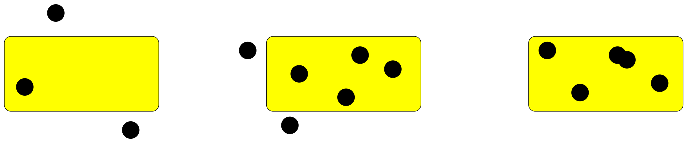


Figure 6: Three Simple Geo-Filtering Predicates: ANY-INSIDE (left), MOST-INSIDE (middle) and ALL-INSIDE (right).

can then filter out documents or parts of documents that do not fall within our geographic area of interest. Given a polygon P described in a query, and a set of locations $L = \ell_1 \ldots \ell_N$ mentioned in a document. Be Δ_i an N-dimensional vector of geographic distances on the geoid between the N locations in a text document d (mentioned with absolute frequencies f_i) and the centroid of P. Then we can use a *filter predicate* $\textsc{Geo-Filter}(f, \Delta)$ to eliminate the document if its spatial "aboutness" is not high enough:

$$\textsc{Score}'(d,P) = \begin{cases} \textsc{Score}(d) & \textsc{Geo-Filter}(f_d, \Delta_d, P) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

In filtering the decision is simply between passing through the original IR score or setting it to 0, thus effectively discarding the document from the ranking. Here are the definitions of three simple GEO-FILTER predicates:

1. ANY-INSIDE. This filter is most conservative and tries to avoid discarding true positives at the risk of under-utilizing the discriminative power of geographic space for IR. It only filters out documents that mention no location in the query polygon P:

$$\text{ANY-INSIDE}(f_d, \Delta_d, P) = \begin{cases} true & \exists_{\ell \in d} : \ell \in P \\ false & \text{otherwise} \end{cases} \tag{5}$$

2. MOST-INSIDE. This filter is slightly more agressive than ANY-INSIDE, but still allows for some noise (locations mentioned that do not fall into the geographic area of interest as described by the query polygon P). It discards all documents that mention more locations that fall outside the query polygon than inside:

$$\text{MOST-INSIDE}(f_d, \Delta_d, P) = \begin{cases} true & |\{\ell \in d | \ell \in P\}| > |\{\ell \in d | \ell \notin P\}| \\ false & \text{otherwise} \end{cases} \tag{6}$$

3. ALL-INSIDE. This filter is perhaps too agressive for most purposes; it discards all documents that mention even a single location that fall outside the query polygon P, i.e. all locations must be in the geographic space under consideration:

$$\text{ALL-INSIDE}(f_d, \Delta_d, P) = \begin{cases} true & \forall_{\ell \in d} : \ell \in P \\ false & \text{otherwise} \end{cases} \tag{7}$$

Figure 6 summarizes the semantics of the three geo-filtering predicates used.

In practice, we use Minimal Bounding Rectangles (MBRs) to approximate the polygons described by the locations in the query, which trades runtime performance against retrieval performance. More specifically, we computed the union of the Alexandria Digital Library and ESRI gazetteers (Table 2) to look up MBRs for geographic terms in the GEO-CLEF queries.[5] In cases of multiple candidate referents (e.g. for *California*), the MBR for the largest feature type was chosen (i.e. in the case of California, the U.S. membership state interpretation). Latin America was not found in the Alexandria Gazetteer. A

[5] On the query side, manual disambiguation was performed.

Expression	Alexandria MBR	ESRI MBR
Asia	(0; 0), (90; 180)	—
Australia	(-45.73; 111.22), (-8.88; 155.72)	(-47.5; 92.2), (10.8; 179.9)
Europe	(35.0; -30.0), (70.0; 50.0)	(35.3; -11.5), (81.4; 43.2)
Latin America	—	(-55.4; -117), (32.7; -33.8)
Bosnia-Herzegovina	(42.38; 15.76), (45.45; 20.02)	—
Germany	(46.86; 5.68), (55.41; 15.68)	(47.27; 5.86), (55.057; 15.03)
Holland	(50.56; 3.54), (53.59; 7.62)	(51.29; 5.08), (51.44; 5.23)
Japan	(30.1; 128.74), (46.26; 146.46)	(24.25; 123.68), (45.49; 145.81)
Rwanda	(-3.01; 28.9), (-1.03; 31.2)	(-2.83; 28.85), (-1.05; 30.89)
UK	(49.49; -8.41), (59.07; 2.39)	(49.96; -8.17), (60.84; 1.75)
United States	(13.71; -177.1), (76.63; -61.48)	(18.93; -178.22), (71.35;-68)
California	(32.02; -124.9), (42.51; -113.61)	—
Scotland	— (56.0; -4.0)	(54.63; -8.62), (60.84; -0.76)
Siberia	— (60.0; 100.0)	—
Scottish Islands	—	—
Scottish Trossachs	— (49.63; -104.22)	—
Scottish Highlands	— (57.5; -4.5)	—
Sarajevo	— (43.86; 18.39)	(43.65; 18.18), (44.05; 18.58)
Caspian Sea	— (42.0; 50.0)	(45; 48.41), (42.40; 48.81)
North Sea	— (55.33; 3.0)	(58.04; 1.02), (58.44; 1.42)

Table 2: Minimal bounding rectangles (MBRs) from the Alexandria and ESRI gazetteers. MBRs are given as pairs of points, each with lat/long in degrees. A dash means that no result was found or that a centroid point was available only.

manual search for South America also did not retrieve the continent, but found several other hits, e.g. South America Island in Alaska. Holland was recognized by the Alexandria Gazetteer as a synonym for the Netherlands. While this corresponds to typical usage, formally speaking Holland refers to a *part* of the Netherlands. The ESRI server returned two entries for *Caspian Sea*, one as given in the table, another with MBR (41.81; 50.54), (42.21; 50.94)–since they share the same feature type they could not otherwise be distinguished. Finally, the software module CLIP performs geographic filtering of a document given an MBR, very much like the clipping operation found in typical GIS packages, albeit on unstructured documents.

It would of course have been beneficial for the retrieval performance if the MBRs that were not available in the ESRI and Alexandria gazetteers had been gathered from elsewhere, as there are plenty of sources scattered across the Internet. However, then the experimental outcome would perhaps no longer reflect a typical *automatic* system.

2.5 Query Expansion with Meronyms

Query expansion is typically used as a Recall-enhancing device, because by adding terms to the original query that are related to the original terms, additional relevant documents are retrieved that would not have been covered by the original query, possibly at the expense of Precision. Here, we experimented with meronym query expansion, i.e. with geographic terms that stand in a spatial “part-of” relation (as in “Germany is part of Europe”). We used WordNet 2.0 to retrieve toponyms that stand in a meronym relationship with any geographic term from the query. The choice of WordNet was motivated by the excessive size of both gazetteers used in the toponym resolution step. For each query, we transitively added all constituent geographic entites, e.g. for *California* we added *Orange County* as well as *Los Angeles*.[6] Figure 7 shows the number of terms that are added for each query. For queries 2, 6, 7, 8, 12 and 19 the number of meronyms in WordNet was actually higher than 1,000; however, an analysis revealed that an

[6] Apostrophies (“’”) were eliminated for technical reasons.

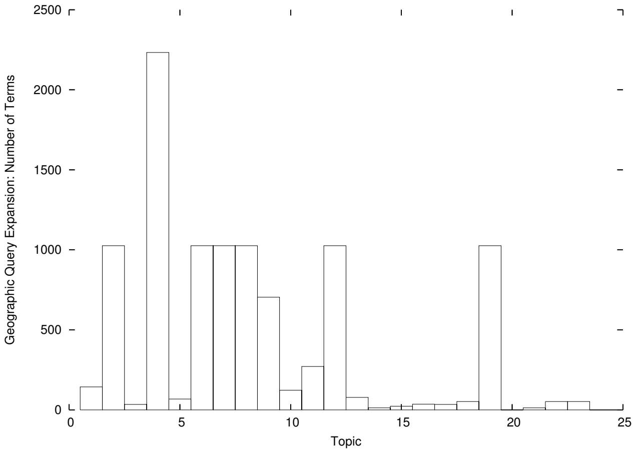


Figure 7: Geographic Query Expansion with Meronyms from WordNet.

implementation limit of *Lucene* was hit that caused a cutoff after 1,000 terms.

The next section describes the evaluation results. Please consult Appendix A for a description of each run.

3 Evaluation

3.1 Method

The GEO-CLEF 2005 evaluation was very similar to previous TREC and CLEF evaluations: for each run, *11-Point-Average Precision* against *interpolated Recall* and *R-Precision* against retrieved documents are plotted. In addition, difference from median across participants for each topic is reported.

Traditionally, the relevance judgments in TREC-style evaluations are binary, i.e. a document either meets the information need expressed in a TREC topic (1) or not (0). Intrinsically fuzzy queries (e.g. "*shark attacks near Australia*") introduce the problem that a strict yes/no decision might no longer be appropriate; there is no "crisp cut-off point. In the same way that the ranking has to be modified to account for geographic distance, a modification of the evaluation procedure ought to be considered. However, for GEO-CLEF 2005, binary relevance assessments were used.

Nota Bene. For organizational reasons, this series of experiments did *not* contribute any documents to the judgment pool for the relevance assessments, which results in a negative bias of the performance results measured compared to the true performance of the experiments and other GEO-CLEF 2005 participants. This is because all relevant documents found by the methods described herein but not returned by any other participants will be have been wrongly assessed as "not relevant". Therefore, a discussion of the relative performance compared to other participants is not included in this paper. On the other hand, this makes the results comparable to future experiments with GEO-CLEF data outside the annual evaluation, which will of course likewise not be able to influence the pooling a posteriori.

We describe the results obtained and present (a very preliminary) discussion.

3.2 Results

Retrieval Performance. In our experiments, the baseline run `LTITLE` that uses only the topic title and no spatial processing performs surprisingly well (Figure 8), with an Average Precision averaged over queries

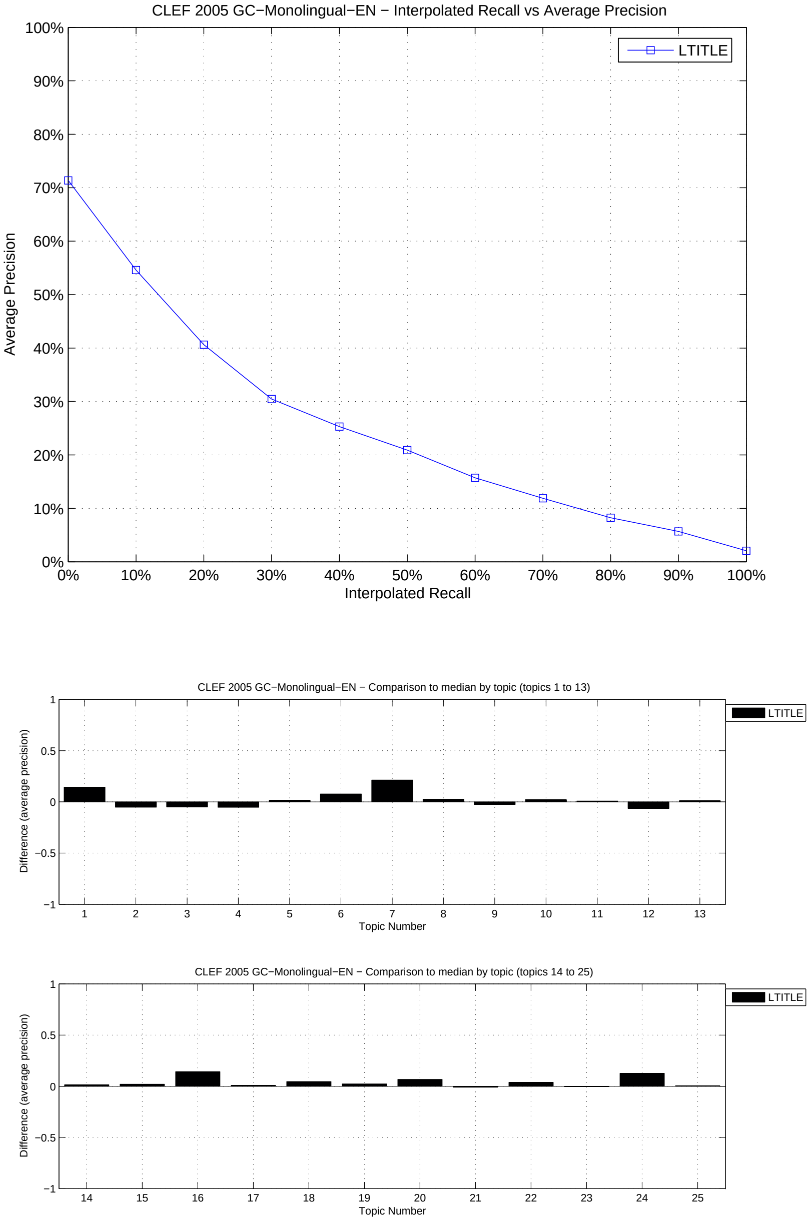


Figure 8: The automatic topic title run (`LTITLE`): Average Precision (left) and performance relative to the median across participants (right).

Run	Avg. Precision	R-Precision
`LTITLE`	**23.62** %	**26.21** %
`LTITLEANY`	**18.50** %	**21.08** %
`LTITLEMOST`	12.64 %	16.77 %
`LTITLEALL`	8.48 %	11.97 %
`LCONCPHR`	15.65 %	19.25 %
`LCONCPHRANY`	14.18 %	19.66 %
`LCONCPHRMOST`	9.56 %	14.46 %
`LCONCPHRALL`	7.36 %	10.98 %
`LCONCPHRSPAT`	**20.37** %	**24.53** %
`LCONCPHRSPATANY`	16.92 %	20.36 %
`LCONCPHRSPATMOST`	11.09 %	15.51 %
`LCONCPHRSPATALL`	7.99 %	10.89 %
`LCONCPHRWNMN`	17.25 %	19.36 %
`LCONCPHRWNMNANY`	12.99 %	16.22 %
`LCONCPHRWNMNMOST`	8.18 %	11.38 %
`LCONCPHRWNMNALL`	5.69 %	8.78 %

Table 3: GEO-CLEF 2005 result summary.

of 23.62% and a Precision at 10 documents of just 36%.

Table 3 gives a summary of the averaged results for each run. As for the terminology, all run names start the letter `L` followed by an indicator of how the query was formed. `CONC` means using the content of the `<CONCEPT>` tag and posing a phrasal query to the IR engine, `CONCPHRSPAT` means using the content of both `<CONCEPT>` and `<SPATIAL>` tags, and `<TITLE>` uses the title tag. `PHR` refers to runs using the IR engine's phrasal query mechanism rather than bag-of-terms. For these runs, queries look as follows:

```
( ("Shark Attacks"ˆ2.0)
  (("shark attack"˜8)ˆ1.5)
  (Shark Attacks) )
```

This combined way of querying takes into account the phrase *shark attacks* (as subsequent terms in the document only) with twice the weight of the "normal" bag-of-words query (last sub-query). The middle line searches for the lemmatized words *shark* and *attack* within an 8-term window and weights this sub-query with 1.5. Runs containing `ANY`, `MOST`, or `ALL` as part of their name indicate that geo-filtering with the ANY-INSIDE, MOST-INSIDE or ALL-INSIDE filtering predicates, respectively, was used. Finally, `WN` as part of a run name indicates that query expansion with WordNet meronyms was applied. Appendix A contains an description of the meaning of the run names.

Runtime Performance. The eximental setup for this study was not optimized for runtime performance. The indexing of the GEO-CLEF document collection took 38 minutes (100 minutes wallclock time on a single machine with Network File System). The execution time for the title-only baseline (run `LTITLE`) was 12 s (runtime averaged over 3 runs).[7] The most expensive operation was the named entity taggin of the result document pool 3,549 min (ca. 60 hours). Gazetteer lookup amounted to 400 min for all pooled result types, and toponym resolution time took 7:30 min, again for all pooled result types. The re-ranking by geographic filtering itself was fast and took only 8 seconds per run.

3.3 Discussion

Applying the "maximum population" heuristic alone to achieve toponym resolution together with geo-filtering in general performed poorly and in none of the four series of experiments outperformed a baseline that applied no dedicated spatial processing.

[7] Runtime performance is based on a 1-CPU Fujitsu-Siemens SCENIC W600-i865G with Intel Pentium 4 processor (2994 MHz, 1 MB cache, 5931 BogoMIPS) running Novel SuSE Linux 9.1 kernel version 2.6.8-24-smp.

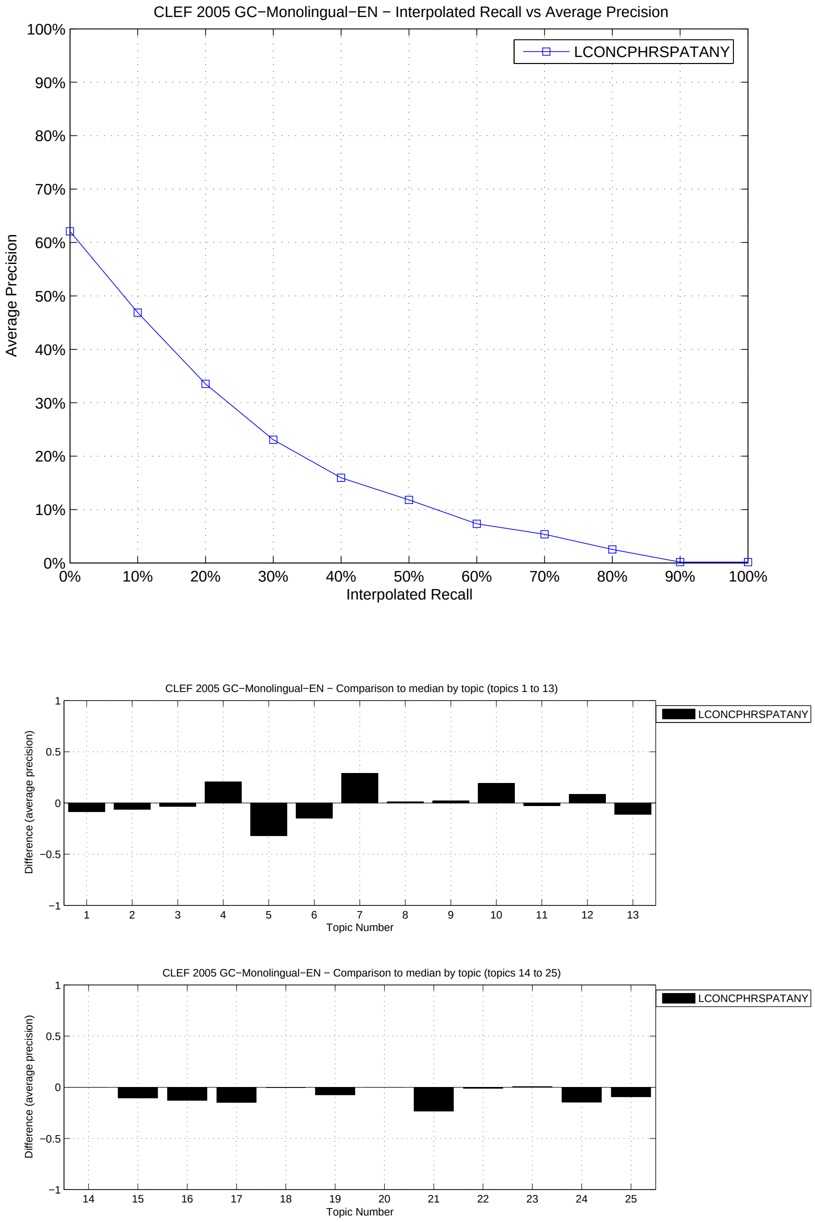


Figure 9: The run `LCONCPHRSPATANY` shows that robustness across queries is an issue.

Interestingly, a plain vanilla Vector Space Model with TF-IDF and the obligatory run using title-only queries (`LTITLE`) performs better than the median across all participant entries for 19 out of 25 (or 76%) of the queries in GEO-CLEF 2005.

For three geo-filtering predicates tested, a consistent relative pattern could be observed across all runs: The ANY-INSIDE filter almost consistently outperformed (in one case it was en par with) the MOST-INSIDE filter, which in turn always outperformed the ALL-INSIDE filter.
While it was expected that MOST-INSIDE would not perform all well as the other two filter types, it is interesting that the conservative ANY-INSIDE outperformed MOST-INSIDE on average.

The evidence seems to suggest further than geographic query expansion with WordNet meronyms is not effective as a recall-enhancing device, independent on whether or which geo-filter is applied afterwards: average precision at. Note however, that this is true only on average, not for all individual queries. Furthermore two queries were actually not executed by the Lucene engine because the query expansion caused the query to exceed implementation limits (too many query terms).

Geo-CLEF Methodology. Regarding the *modus operandi* of GEO-CLEF, future evaluations would benefit from a separation of training/development and test set regarding the queries.

Furthermore, alternative relevance assessments based on geographic distance rather than binary decisions (document relevant/document not relevant) might be attempted. For instance, *Root-Mean-Square Distance* (RMSD, Equation 8) could be used to indicate the (geo-)distance between a query centroid q and a set of location centroids $d_1, \ldots, d_N$ in a document:

$$\mathrm{RMSD}(d,q) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - q)^2} \tag{8}$$

Such an measure could be used to compute a continuous-scale geographic relevance measure once the assessors annotated the test queries and the toponyms in the pooled result documents with their “ground truth” coordinates.

Geo-CLEF Dataset. To better understand the low performance of the experiment, we performed a manual analysis of topic `GC0001`. To find a large pool of potentially relevant documents, we retrieved the results for the queries

```
(shark OR sharks) AND (attack OR attacked OR attacks)
(shark OR sharks) AND (kill OR killed OR kills)
```

i.e. we are looking for documents in which sharks get mentioned together with the verbs attack (61 hits) or kill (63 hits), respectively, no matter where the potential attack event described happens. Since *shark* does not have well-known synonyms, and since we use several forms of *to attack* and the even stronger *to kill* (in case the attack itself is not focused on but rather its outcome), we expect these two queries to cover most relevant documents for the first topic (together resulting in a pool of size 107 documents). The aim of our microscopic analysis is to find out whether the mechanisms applied are at all meaningful, given the dataset. For example, the GIR method proposed in this paper could be worthless if all mentions of *Perth*, actually conincided with mentions of *Australia*, because then the query term `Australia` would then capture relevant documents directly. Indeed, in document `GH951219-000021`, we find

```
PERTH : A severed human arm wrapped in a torn piece of wetsuit has
washed up on a beach more than three months after a shark killed a
29-year-old scuba diver , police in Australia said .
```

We went through the retrieved document set and carried out a relevance assessment, bearing in mind the geography.[8] Table 4 shows the result. Only 11 (or less than 10.28%) of retrieved documents are actually about shark attack events, of which 4 contain *California* and 1 contains *Australia*. Only one document was dealing with a story outside the geographic scopy of the first GEO-CLEF query (in bold type), whereas

[8] Funnily enough, documents `LA030994-0075`, `LA053094-0133`, `LA121594-0181` and `LA121594-0267` contain stories of type “man bites dog”: they report about a large initiative of people killing sharks, not vice versa, which means we have to judge them *not relevant*. We do not count `LA103094-0316` as relevant, which reports that former Australian prime minister Holt was *believed* to have been eaten by sharks, since it is not actually a report on an established shark attack event.

Document ID	Query Term Mentioned	Other Toponyms Mentioned
“attack” query		
LA041794-0356	California	Point Loma, San Diego
LA051994-0068	California	El Toro, Guadalupe Island
LA122194-0180	California	Santa Barbara
“kill” query		
GH950614-000127	**—**	**Clearwater Bay, Hong Kong**
GH951219-000021	Australia	Perth
LA010294-0151	—	Seattle, San Diego, Buffalo, Minneapolis, St. Thomas, Virgin
LA030994-0082	—	Hawaii, Maui, Maile Point, Oahu, Lahaina, Maui, Honolulu
LA072894-0159	—	Orange County, America, Santa Cruz, Manly, Australia, Java, Durban, Jeffries Bay, South Africa, Johannesburg, London, Los Angeles, Las Vegas Islands, New Jersey, Florida, Hawaii
LA121594-0132	—	Los Angeles, Santa Barbara County
both “kill” and “attack” occur		
LA121094-0160	California	Northern California, San Diego, San Miguel Island, Santa Barbara
LA041994-0146	—	San Diego, Point Loma, United States

Table 4: Documents About Shark Attack Events.

5 documents report about shark attacks in Calfornia while not explicitly mentioning California; these are the interesting cases where geo-filtering or any other dedicated GIR technique could have helped, and this statistics shows that it would have helped recall rather than precision since more relevant documents unretrievable by the geographic search terms in the query would have been retrieved in addition than non-relevant documents excluded on grounds of geographic irrelevance. Geo-filtering as proposed here could achieve this recall-enhancing function, most likely because of the limited population gazetteer used here, and its precision-enhancing ability could not be demonstrated on the GEO-CLEF 2005 corpus, perhaps because of this unfortunate ratio between out-of-geographic-scope documents of 1/107 at least not for this first query studied.

Another problem is that Schotland and California are both used for corpus sampling and as regions in the query. Since many articles contain the place of publication in headers or footers. Since in these experiments, no dedicated position-dependent document analysis was carried out this could have introduced noise. The fact that the second half of the queries performs much lower is due to the fact that despite merging two gazetteer sources, our gazetteer used is still not dense enough to cover e.g. the Scottish Trossachs, the Scottish Highlands or even Siberia. Finally, the number of queries in GEO-CLEF 2005 was quite small (only 25 queries). As a result, the problems mentioned before can in combination easily overshadow any algorithm’s performance, which a more detailed analysis would have to show.

4 Conclusions and Future Work

4.1 Conclusions

We have described a method for geographic information retrieval based on named entity tagging to identify place names (or toponym recognition, geo-parsing), toponym resolution (or geo-coding, place name disambiguation) and geographic filtering (or clipping).

First results show that a very simple method for toponym resolution based on a “maximum population” heuristic is not effective when combined with three point-in-MBR geo-filtering predicates in the setting used. We conjecture this may be due to the lack of available population data. In addition, we discovered that geographic query expansion with WordNet meronyms appears not to improve retrieval performance. However, a deeper analysis of the results will be necessary before drawing any definite conclusions.

4.2 Future Work

For future work, several opportunities for further study should be given consideration:

1. The results presented here should be compared the with different, more sophisticated clipping criteria that take the amount of spatial overlap into account. For example, instead of using MBRs computed from sets of centroid points [AJT01] proposes a *Dynamic Spatial Approximation Method (DSAM)*, which uses Vonoroi approximation to compute more precise polygons from sets of points. Once polygons are available, spatial overlap metrics can be applied to improve retrieval [RF04].

2. It is vital to discover methods to determine a good balance when weighting the spatial influence and the term influence in the query against each other in a principled way, probably even dependent on the query type.

3. On the query side, the specific spatial relations should be taking into account. However, this requires defining how users and/or CLEF assessors actually judge different relations beforehand (how near does something have to be to be considered "near"?).

4. On the document side, text-local relationships from the toponym context should be taken into account. Right now, all toponyms (`LOC`) are considered equal, which does not utilize knowledge from the context of their occurrence. For instance, a document collection that has one mention of *New York* in every document footer because the news agency resides in New York can pose a problem.

5. The impact of the particular gazetteer used for query expansion and toponym resolution ought to be studied with respect to the dimensions size/density (UN-LOCODE/WordNet versuss NGA GeoNames) and local/global (e.g. EDINA DIGIMAP versus NGA GeoNames).

6. Last but perhaps most importantly, more sophisticated toponym resolution strategies (e.g. [LSW03]) should be compared against the simple population heuristic used in this study.

Acknowledgements

The author is grateful to Prof. Dr. Dietrich Klakow for discussions of the results and for hosting him in his lab during a visit in 2005. This work was supported in part by the SOCRATES program of the European Union.

References

[AJT01] Harith Alani, Christopher B. Jones, and Douglas Tudhope. Voronoi-based region approximation for geographical information retrieval with gazetteers. *International Journal of Geographical Information Science*, 15(4):287–306, 2001.

[Av04] Avi Arampatzis and Marc van Kreveld. Practical similarity ranking. EU Project Deliverable, Spatially-Aware Information Retrieval on the Internet, IST-2001-35047 D18:5302, 2004.

[CC03] James R. Curran and Stephen Clark. Language independent NER using a maximum entropy tagger. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-03)*, pages 164–167, Edmonton, Canada, 2003.

[Cut05] Doug Cutting. Lucene. http://lucene.apache.org/, 2005. [online].

[FJA05] F. Fu, C. B. Jones, and A. I. Abdelmoty. Building a geographical ontology for intelligent spatial search on the Web. In *Proceedings of IASTED International Conference on Databases and Applications (DBA-2005)*, Innsbruck, Austria, 2005. IASTED.

[GH05] Otis Gospodnetić and Erik Hatcher. *Lucene in Action*. Manning, Greenwich, CT, USA, 2005.

[Lei] Jochen L. Leidner. An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems. Special Issue on Geographic Information Retrieval*. (in press).

[LSW03] Jochen L. Leidner, Gail Sinclair, and Bonnie Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the Workshop on the Analysis of Geographic References held at the Joint Conference for Human Language Technology and the Annual Meeting of the Noth American Chapter of the Association for Computational Linguistics 2003 (HLT/NAACL'03)*, pages 31–38, Edmonton, Alberta, Canada, 2003.

[RF04] Larson Ray R. and Patricia Frontiera. Spatial ranking methods for geographic information retrieval (GIR) in digital libraries. In *Research and Advanced Technology for Digital Libraries, 8th European Conference, ECDL 2004, Bath, UK, September 12-17, 2004, Proceedings*, volume 3232 of *Lecture Notes in Computer Science*, pages 45–56. Springer, 2004.

[SC01] David A. Smith and Gregory Crane. Disambiguating geographic names in a historical digital library. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '01)*, pages 127–136, London, UK, 2001. Springer-Verlag.

[van05] Ben van Klinken. CLucene. http://sourceforge.net/projects/clucene/, 2005. [online].

[vOV91] P. van Oosterom and T. Vijlbrief. Building a GIS on top of the open DBMS Postgres. In *Proceedings of the European Conference on Geographic Information Systems (EGIS)*, pages 775–787, 1991.

[vRAvZ05] Marc van Kreveld, Iris Reinbacher, Avi Arampatzis, and Roelof van Zwol. Multi-dimensional scattered ranking methods for geographic information retrieval. *GeoInformatica*, 9(1):61–84, 2005.

A Descriptions of Runs

Name of GEO-CLEF 2005 run	Description
LCONCPHR	concept phrasal query
LCONCPHRALL	concept phrasal query toponym resolution based on population heuristic all-inside geo-filtering
LCONCPHRANY	concept phrasal query toponym resolution based on population heuristic any-inside geo-filtering
LCONCPHRMOST	concept phrasal query toponym resolution based on population heuristic most-inside geo-filtering
LCONCPHRSPAT	concept phrasal query with spatial aspect
LCONCPHRSPATALL	concept phrasal query with spatial aspect toponym resolution based on population heuristic all-inside geo-filtering
LCONCPHRSPATANY	concept phrasal query with spatial aspect toponym resolution based on population heuristic any-inside geo-filtering
LCONCPHRSPATMOST	concept phrasal query with spatial aspect toponym resolution based on population heuristic most-inside geo-filtering
LCONCPHRWNMN	concept phrasal query with spatial aspect WordNet meronym query expansion
LCONCPHRWNMNALL	concept phrasal query with spatial aspect WordNet meronym query expansion toponym resolution based on population heuristic all-inside geo-filtering
LCONCPHRWNMNANY	concept phrasal query with spatial aspect WordNet meronym query expansion toponym resolution based on population heuristic any-inside geo-filtering
LCONCPHRWNMNMOST	concept phrasal query with spatial aspect WordNet meronym query expansion toponym resolution based on population heuristic most-inside geo-filtering
LTITLE	title query
LTITLEALL	title query toponym resolution based on population heuristic all-inside geo-filtering
LTITLEANY	title query toponym resolution based on population heuristic any-inside geo-filtering
LTITLEMOST	title query toponym resolution based on population heuristic most-inside geo-filtering